Inconsistencies in Estimating the Age of HIV-1 Subtypes Due to Heterotachy

Joel O. Wertheim,[1,*] Mathieu Fourment,[1,†] and Sergei L. Kosakovsky Pond[2]

[1]Department of Pathology, University of California, San Diego

[2]Department of Medicine, University of California, San Diego

†Present address: Program of Emerging Infectious Diseases, Duke-NUS Graduate Medical School, Singapore

***Corresponding author:** E-mail: jwertheim@ucsd.edu.

Associate editor: Sudhir Kumar

Abstract

Rate heterogeneity among lineages is a common feature of molecular evolution, and it has long impeded our ability to accurately estimate the age of evolutionary divergence events. The development of relaxed molecular clocks, which model variable substitution rates among lineages, was intended to rectify this problem. Major subtypes of pandemic HIV-1 group M are thought to exemplify closely related lineages with different substitution rates. Here, we report that inferring the time of most recent common ancestor of all these subtypes in a single phylogeny under a single (relaxed) molecular clock produces significantly different dates for many of the subtypes than does analysis of each subtype on its own. We explore various methods to ameliorate this problem. We conclude that current molecular dating methods are inadequate for dealing with this type of substitution rate variation in HIV-1. Through simulation, we show that heterotachy causes root ages to be overestimated.

Key words: molecular clock, rate variation, HIV-1.

Appropriate modeling of substitution rate variation among lineages, heterotachy, is critical to many problems in evolutionary biology, especially molecular dating (Kumar 2005). Among all approaches that permit lineage-specific rate variation, one method has recently risen to prominence when inferring substitution rates and times of most recent common ancestors (tMRCAs): applying a distribution of uncorrelated substitution rates across a phylogenetic tree. However, whether these models perform adequately when substitution rates differ dramatically among clades has recently been called into question (Ho and Lanfear 2010; Dornburg et al. 2011).

There are nine HIV-1 group M subtypes and over 40 circulating recombinant forms (Hemelaar et al. 2006), and recent estimates place the tMRCA of HIV-1 M in 1908 (1884–1924) (Worobey et al. 2008). Substitution rates vary extensively among HIV-1 group M subtypes (Penn et al. 2008; Abecasis et al. 2009), yet the impact of heterotachy on tMRCA inference remains underappreciated. In their investigation into the emergence of nosocomial subtype F in Romania, Mehta et al. (2011) noted that phylogenetic inference performed on only subtype F sequences produced tMRCAs consistent with the historical record, whereas inference performed with subtype F and additional HIV-1 M sequences yielded substantially older counterfactual tMRCAs. A similar inconsistency can be seen in dating the emergence of subtype B in the Americas, which placed the tMRCA of subtype B in 1966 (1962–1970) (Gilbert et al. 2007); however, molecular dating using an HIV-1 M data set indicated that the tMRCA of subtype B existed well before 1960 (Worobey et al. 2008). This discrepancy is equivocal, as these two studies analyzed different genomic regions from different viral isolates; yet, it raises a question of sensitivity to sampling.

Here, we investigated the impact of among-lineage rate variation when inferring the age of HIV-1 M and five of its major, nonrecombinant subtypes: A1/A2, B, C, D, and F1/F2, chosen because there are sufficient numbers of nucleotide sequences available in public databases for robust phylogenetic inference. We estimated the tMRCA for each subtype using isolates from only a given subtype and also from the entire HIV-1 M phylogeny (i.e., combined HIV-1 M analysis, see Methods).

Based on an uncorrelated lognormally distributed relaxed molecular clock model (Drummond et al. 2006), we corroborated the finding by Abecasis et al. (2009) that mean substitution rate estimates differ as much as 2-fold between subtypes. Subtypes C and D had similar substitution rates to those inferred from the combined HIV-1 M analysis, whereas A1/A2, B, and F1/F2 all had significantly faster subtype-specific rates ($P > 0.95$, fig. 1*a*). Phylogenetic relatedness of subtypes did not translate into correspondingly similar rates (fig. 1*b*), akin to a previous finding with codon-substitution models (Pond et al. 2010). There were also appreciable differences in the tMRCAs inferred from subtype-only analyses and inference using the combined HIV-1 M data set (fig. 2 and table 1). Subtypes B and F1/F2-specific tMRCAs were younger than those from the combined analysis ($P > 0.95$); subtype A1/A2 analysis suggested a marginally significant younger tMRCA for A1 ($P = 0.949$). D was the sole subtype whose tMRCA was older in the subtype-only inference than in the combined analysis. The inferred tMRCA of HIV-1 M was in agreement with

Letter

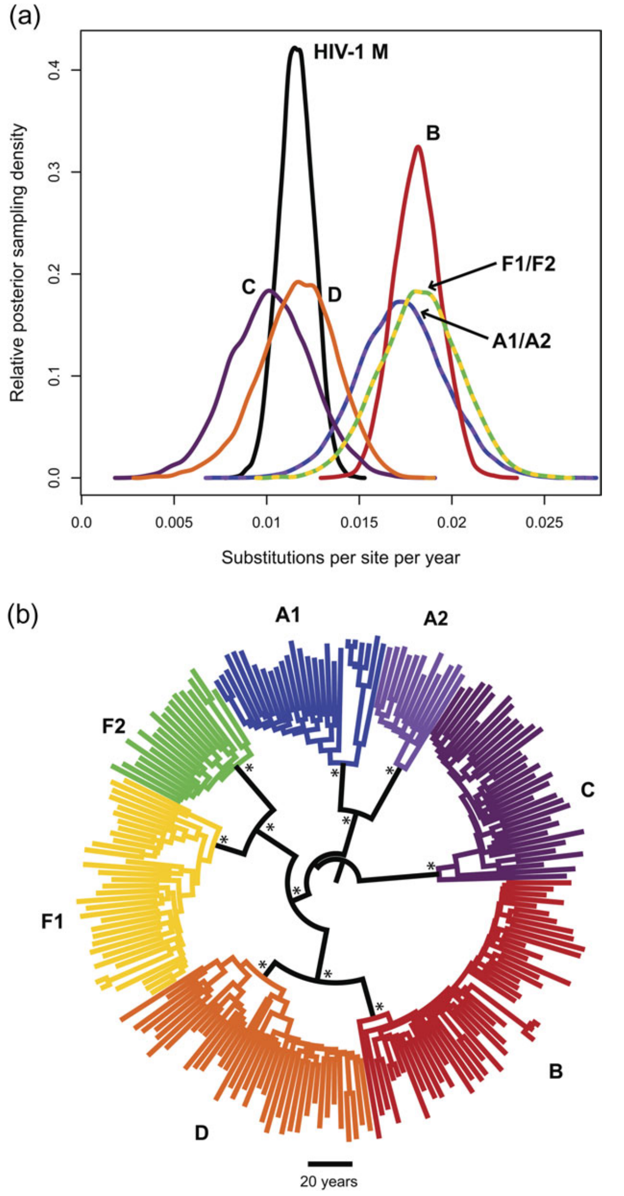

Fig. 1. Substitution rates vary across the HIV-1 M phylogeny. (*a*) Posterior distributions of mean substitution rates for subtype-only and combined HIV-1 M analyses. (*b*) Maximum clade credibility tree for HIV-1 M subtypes A1/A2, B, C, D, and F1/F2. Highly supported nodes (posterior probability $>$ 0.9) differentiating relationships among subtypes are indicated with asterisks.

previous studies (Worobey et al. 2008; Wertheim and Worobey 2009). We hypothesize that the differences in tMRCA estimates between the subtype-only and combined HIV-1 M analyses resulted from the inability of the relaxed molecular clock to capture the extensive substitution rate variation likely present among the subtypes.

The discrepancy between subtype-specific and HIV-1 M analyses was not restricted to subtype tMRCAs. Major lineages within subtypes (F1, F2, A1, and A2) also had different tMRCAs between analyses (fig. 2 and table 1). Within subtype F1, the tMRCA of epidemics in Romania and Brazil varied dramatically between the subtype-only and combined HIV-1 M analyses (as seen in Mehta et al. 2011).

We explored avenues to resolve the discrepancies between the subtype-only and combined HIV-1 M tMRCA estimates. Ho and Lanfear's (2010) suggestion of modeling rate variation among lineages by partitioning the molecular clock so that each codon position (first, second, and third) has its own relaxed clock produced a substantially better fit [$\log_{10}$ Bayes factor $=$ 59.4], but the inferred subtype tMRCAs changed very little (table 1).

The aforementioned analyses employed a Bayesian skyline plot coalescent prior, which has the fewest demographic constraints (Drummond et al. 2005). Fitting alternate, more-restrictive coalescent models (e.g., constant size, expansion growth, exponential growth, and lognormal growth) produced different tMRCA estimates (supplementary table S1, Supplementary Material online); similar results were reported by Worobey et al. (2008). Inference using the alternate models generally estimated younger tMRCAs for the subtypes and the root, except for the expansion growth model. However, using alternate coalescent models did not resolve the discrepancy between the subtype-only and combined analyses. Therefore, we fit a model in which each subtype had its own independent exponential growth rate (for a discussion of this approach, see Ho et al. 2008; Bjork et al. 2011). The tMRCAs inferred using this hybrid model, with multiple coalescent priors, were more similar to the combined HIV-1 M analysis under a Bayesian skyline plot than the subtype-only analyses (table 1).

Abecasis et al. (2009) hypothesized that the variable substitution rates among HIV-1 M subtypes may be due to different selective pressures operating on them. When we inferred the tMRCA of the combined HIV-1 M phylogeny using a codon model accounting for differing selection pressures among sites (Goldman and Yang 1994; Suchard and Rambaut 2009), the inferred tMRCAs were indistinguishable from those of the more simple general time reversible (GTR) $+ \Gamma_4$ nucleotide model (table 1). This is not surprising, given a recent report that modeling lineage-specific rate variation has a much more pronounced effect, than modeling codon substitutions, on tMRCA estimates (Wertheim and Kosakovsky Pond 2011).

Next, we explored the utility of local molecular clocks (Drummond and Suchard 2010) in resolving the discrepancy between tMRCA estimates. The local clock detected 7.6 (5–10) rate changes throughout the HIV-1 M phylogeny, though none of these changes occurred near the base of the subtypes. A slightly faster mean rate was inferred under the local clock: 1.29×10^{-3} (1.15×10^{-3}–1.15×10^{-3}) substitutions/site/year compared with 1.16×10^{-3} (9.81×10^{-4}–1.37×10^{-3}) substitutions/site/year in the combined HIV-1 M analysis using an uncorrelated lognormally distributed clock. Correspondingly, the tMRCA estimates inferred using a local clock were younger (table 1).

To determine the expected effect of substitution rate variation across HIV-1 M subtypes on tMRCA inference, we simulated nucleotide sequences across the combined HIV-1 M phylogeny by applying the subtype-specific mean substitution rates. These simulations yielded subtype tMRCAs that were even more inconsistent with the

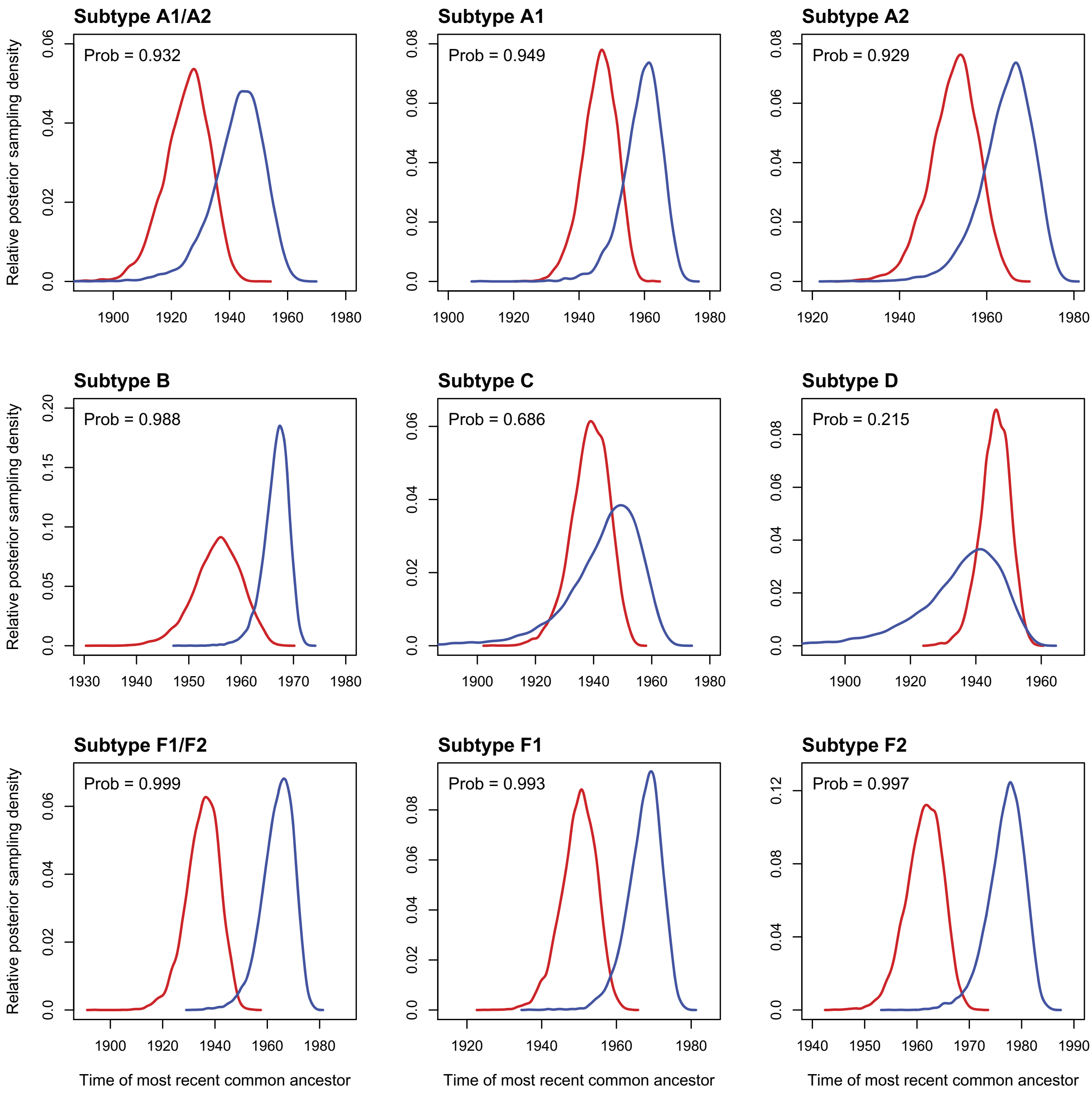

FIG. 2. Posterior distributions of tMRCA estimates for HIV-1 M subtypes. The probabilities that the subtype-only posteriors (blue) are younger than the combined HIV-1 M posteriors (red) are shown.

subtype-only tMRCAs than the combined HIV-1 M analysis (table 1). For example, simulated subtypes A1 and F1 each were 10 years older than in the combined HIV-1 M analysis, whereas subtype D was 13 years younger. These simulations support the hypothesis that if extreme rate variation does exist among HIV-1 subtypes, a single relaxed molecular clock cannot adequately correct for it.

To determine if factors other than heterotachy were responsible for the inconsistent subtype ages, we simulated nucleotide sequence alignments under a single rate (i.e., no heterotachy) across an HIV-1 chronogram containing all subtype sequences analyzed here (578 sequences, see supplementary table S2, Supplementary Material online). The tMRCA of the HIV subtypes was inferred for both subtype-only and combined HIV-1 M data sets. No significant differences between subtype tMRCAs were found, suggesting that heterotachy was responsible for the subtype tMRCA inconsistencies.

Finally, we performed a series of general simulations to assess the effect of heterotachy on tMRCA estimation (fig. 3*a*). When the substitution rate in a single clade was increased relative to the rest of the phylogeny, the tMRCA inferred for that clade was overestimated; if the rate was decreased, the tMRCA for that clade was underestimated (fig. 3*b*). The tMRCAs inferred for other clades in the phylogeny experienced the opposite trend. Somewhat unexpectedly, more distantly related clades were more biased than closely related clades (see clade B vs. clades C/D in

Table 1. tMRCA for HIV-1 M and Its Subtypes (mean and 95% highest posterior density) Inferred under Different Evolutionary Models and Simulations.

Lineage	Subtype-Only	Combined HIV-1 M	Codon Partitions (first, second, and third)	Multiple Coalescent Priors	Codon Model (GY94 + Γ_4)	Local Clock	HIV-1 M Heterotachy Simulations
A1/A2	1943 (1926–1960)	1925 (1910–1940)	1927 (1914–1939)	1924 (1906–1941)	1926 (1910–1940)	1934 (1925–1942)	1910 (1896–1924)
A1	1959 (1947–1970)	1946 (1936–1956)	1948 (1939–1955)	1945 (1934–1957)	1946 (1935–1956)	1952 (1947–1958)	1936 (1926–1946)
A2	1965 (1953–1975)	1952 (1941–1963)	1954 (1945–1963)	1958 (1947–1968)	1952 (1941–1962)	1959 (1952–1965)	1944 (1934–1953)
B	1967 (1962–1971)	1955 (1946–1964)	1957 (1949–1965)	1955 (1947–1963)	1954 (1945–1962)	1959 (1953–1963)	1948 (1941–1955)
C	1943 (1917–1964)	1939 (1926–1951)	1940 (1930–1950)	1946 (1935–1957)	1938 (1925–1950)	1947 (1940–1954)	1953 (1924–1962)
D	1934 (1905–1956)	1946 (1937–1955)	1947 (1940–1954)	1947 (1938–1956)	1946 (1937–1955)	1952 (1946–1956)	1953 (1945–1959)
F1/F2	1964 (1952–1975)	1935 (1923–1947)	1938 (1928–1947)	1936 (1923–1950)	1936 (1923–1948)	1943 (1935–1950)	1925 (1912–1937)
F1	1967 (1958–1976)	1950 (1940–1959)	1952 (1944–1959)	1953 (1943–1962)	1950 (1941–1959)	1955 (1948–1961)	1940 (1930–1948)
F2	1977 (1970–1983)	1961 (1954–1968)	1963 (1957–1968)	1963 (1954–1971)	1962 (1955–1968)	1966 (1961–1970)	1952 (1945–1959)
HIV-1 M	— —	1904 (1885–1919)	1906 (1892–1919)	1902 (1882–1921)	1903 (1885–1920)	1915 (1905–1924)	1900 (1887–1914)

fig. 3*b*). Notably, the root age was biased further back in time in the presence of substantial heterotachy, regardless of whether the substitution rate increased or decreased. Relaxed clock analysis was able to detect rate changes in a single clade, but the magnitude of this change was substantially underestimated (data not shown). When two clades were simulated under a slower substitution rate (halved), the tMRCA of those clades was underestimated, whereas the tMRCA of the other clades was overestimated (fig. 3*c*). When two clades were simulated under faster rates (doubled), the opposite trend was observed. These patterns were manifested irrespective of the phylogenetic relationships of the clades.

None of the approaches investigated here were able to resolve discrepancies in tMRCA estimation between the subtype-only and the combined HIV-1 M analyses. The implementations of relaxed molecular clocks studied here appear unable to handle the rate variation present in HIV-1 M. When dating the emergence of HIV subtypes and other recent lineages, it seems prudent to use only the clades of interest, as these inferences have generally been more consistent with the historical record. Moreover, given our inability to account for the discrepancies between the subtype tMRCAs and the combined HIV-1 M analysis, we feel less confident in the inferred tMRCA of HIV-1 M. And based on our simulations, one might expect the true tMRCA of HIV-1 M to be younger than previously inferred.

Methods

HIV-1 M subtype data sets were constructed from a nonoverlapping polymerase gene region (HXB2 nucleotide positions 2292–5041) from complete genomes with known years of isolation, downloaded from the Los Alamos National Laboratory HIV Sequence Database (www.hiv.lanl.gov; supplementary table S2, Supplementary Material online). Additional A2, F1, and F2 polymerase sequences (>900 nt) were included to ensure robust substitution rate inference. Recombinants were identified using SCUEAL (Kosakovsky Pond et al. 2009) and removed from subsequent analyses. Alignment was trivial due to the lack of insertions and deletions and was performed manually in Se-Al v2.0 (tree.bio.ed.ac.uk/software/seal/). To create a computationally tractable HIV-1 M data set containing all five subtypes of interest, a maximum of two sequences per sampling year per subtype was included. In all cases, we ensured that sequences from the basal lineage inferred in the subtype-only analyses were included. This procedure resulted in an alignment containing 215 sequences (supplementary table S2, Supplementary Material online). As a control, we investigated the effect of including one or three sequences per sampling year per subtype on tMRCA inference. This effect was negligible for three sequences; analyses sampling one sequence had greater variance which trended deeper in time, as would be expected. Data sets are available at www.hyphy.org/wiki/subtyperates.

Phylogenetic and dating analyses were performed using Bayesian Markov chain Monte Carlo implemented in BEAST v1.6.1 (Drummond and Rambaut 2007). For each analysis, two to four chains of 25 or 50 million generations were run; the first 10% of generations were discarded as burn-in, and chains were combined using LogCombiner. Nucleotide-based analyses were performed using a GTR + Γ_4 substitution model. Codon-based phylogenetic inference using a GY94 + Γ_4 substitution model was implemented in BEAGLE (Suchard and Rambaut 2009). Tracer v1.5 was used to check for convergence and adequate mixing (i.e., estimated sample size > 200 for relevant parameters).

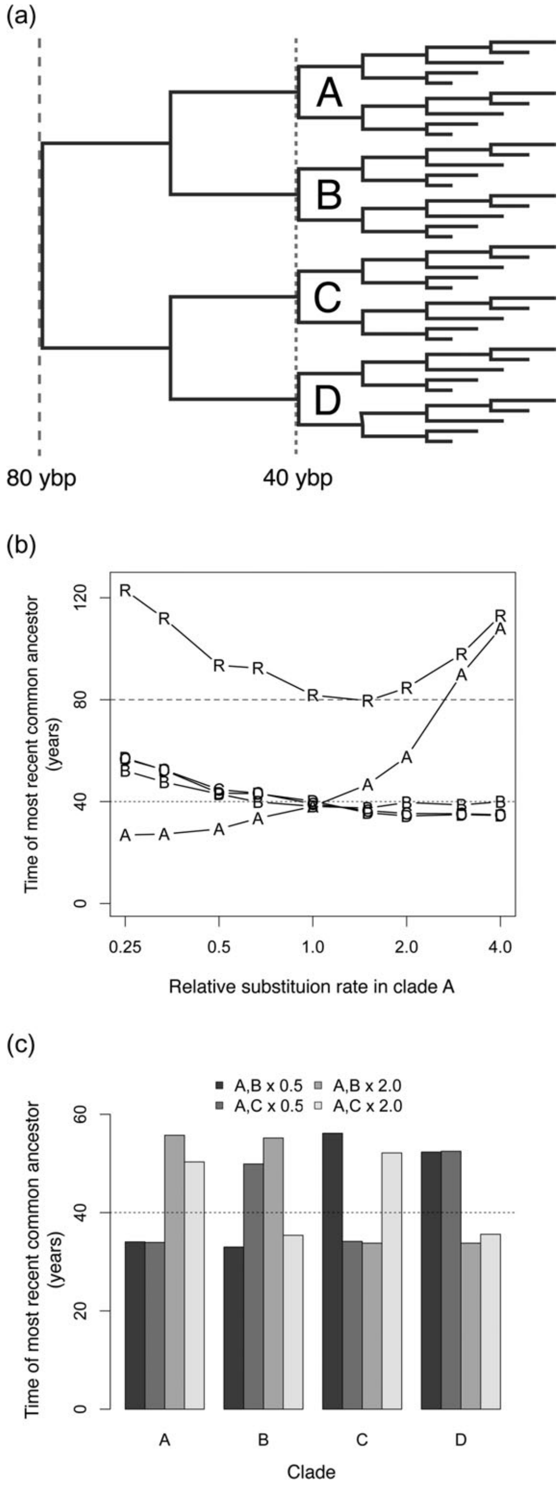

Fig. 3. General heterotachy simulations. (*a*) Chronogram on which simulations were performed. The root of the tree has an age of 80 years before present (ybp), and the four major internal clades have an age of 40 ybp. Clades A, B, and C were permitted to experience increases and decreases in mean substitution rates. (*b*) tMRCA for the root [R] and internal clades (A, B, C, and D) as the mean rate in clade A varies. (*c*) tMRCA for the internal clades when multiple clades were simulated under a slower (0.5×) or faster (2.0×) substitution rate.

HIV-1 heterotachy simulations were performed using the maximum clade credibility tree (chronogram) from the HIV-1 M combined analysis using mean GTR + Γ_4 parameters inferred from the BEAST analysis. Single-rate simulations were performed on a 578-taxon maximum likelihood phylogeny inferred using PhyML v3.0 (Guindon et al. 2010); the chronogram was optimized using TipDate (Rambaut 2000), implemented in HyPhy (Kosakovsky Pond et al. 2005). Nucleotide sequences were simulated using SeqGen v1.3.2 (Rambaut and Grassly 1997). For both sets of HIV simulations, 20 replicate data sets were analyzed in BEAST, using relaxed (lognormal) and strict clocks.

General simulations were performed on a 40-taxon phylogeny comprised of four identical 10-taxon clades (fig. 3*a*). Alignments of 1,000 nt were simulated under a Hasegawa-Kishino-Yano + Γ_4 model. The substitution rate of 1×10^{-3} sites/year was multiplied by a random sample from a lognormal distribution (mean = 0.01; standard deviation = 0.5). To emulate heterotachy, the mean rate within certain clades was varied. For each rate configuration, 50 replicate data sets were analyzed in BEAST.

Supplementary Material

Supplementary tables S1 and S2 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgment

We thank N. Lance Hepler for his assistance with computational resources. This research was supported in part by the National Institutes of Health (AI47745, AI57167, AI74621, and GM093939), the Joint DMS/NIGMS Mathematical Biology Initiative through grant NSF-0714991, the University of California Universitywide AIDS Research Program (grant number IS02-SD-701), a University of California, San Diego Center for AIDS Research/NIAID Developmental Award to S.L.K.P. (AI36214), and a National Institutes of Health Training Fellowship to J.O.W. (AI43638).

References

Abecasis AB, Vandamme AM, Lemey P. 2009. Quantifying differences in the tempo of human immunodeficiency virus type 1 subtype evolution. *J Virol.* 83:12917–12924.

Bjork A, Liu W, Wertheim JO, Hahn BH, Worobey M. 2011. Evolutionary history of chimpanzees inferred from complete mitochondrial genomes. *Mol Biol Evol.* 28:615–623.

Dornburg A, Brandley MC, McGowen MR, Near TJ. 2011. Relaxed clocks and inferences of heterogeneous patterns of nucleotide substitution and divergence time estimates across whales and dolphins (Mammalia:Cetacea). *Mol Biol Evol.* 29:721–736.

Drummond AJ, Ho SY, Phillips MJ, Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. *PLoS Biol.* 4:e88.

Drummond AJ, Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol.* 7:214.

Drummond AJ, Rambaut A, Shapiro B, Pybus OG. 2005. Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol Biol Evol.* 22:1185–1192.

Drummond AJ, Suchard MA. 2010. Bayesian random local clocks, or one rate to rule them all. *BMC Biol.* 8:114.

Gilbert MT, Rambaut A, Wlasiuk G, Spira TJ, Pitchenik AE, Worobey M. 2007. The emergence of HIV/AIDS in the Americas and beyond. *Proc Natl Acad Sci U S A.* 104:18566–18570.

Goldman N, Yang Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol.* 11:725–736.

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol.* 59:307–321.

Hemelaar J, Gouws E, Ghys PD, Osmanov S. 2006. Global and regional distribution of HIV-1 genetic subtypes and recombinants in 2004. *AIDS* 20:W13–W23.

Ho SY, Lanfear R. 2010. Improved characterisation of among-lineage rate variation in cetacean mitogenomes using codon-partitioned relaxed clocks. *Mitochondrial DNA* 21:138–146.

Ho SY, Larson G, Edwards CJ, Heupink TH, Lakin KE, Holland PW, Shapiro B. 2008. Correlating Bayesian date estimates with climatic events and domestication using a bovine case study. *Biol Lett.* 4:370–374.

Kosakovsky Pond SL, Frost SD, Muse SV. 2005. HyPhy: hypothesis testing using phylogenies. *Bioinformatics* 21:676–679.

Kosakovsky Pond SL, Posada D, Stawiski E, Chappey C, Poon AF, Hughes G, Fearnhill E, Gravenor MB, Leigh Brown AJ, Frost SD. 2009. An evolutionary model-based algorithm for accurate phylogenetic breakpoint mapping and subtype prediction in HIV-1. *PLoS Comput Biol.* 5:e1000581.

Kosakovsky Pond SL, Scheffler K, Gravenor MB, Poon AF, Frost SD. 2010. Evolutionary fingerprinting of genes. *Mol Biol Evol.* 27:520–536.

Kumar S. 2005. Molecular clocks: four decades of evolution. *Nat Rev Genet.* 6:654–662.

Mehta SR, Wertheim JO, Delport W, Ene L, Tardei G, Duiculescu D, Kosakovsky Pond SL, Smith DM. 2011. Using phylogeography to characterize the origins of the HIV-1 subtype F epidemic in Romania. *Infect Genet Evol.* 11:975–979.

Penn O, Stern A, Rubinstein ND, Dutheil J, Bacharach E, Galtier N, Pupko T. 2008. Evolutionary modeling of rate shifts reveals specificity determinants in HIV-1 subtypes. *PLoS Comput Biol.* 4:e1000214.

Rambaut A. 2000. Estimating the rate of molecular evolution: incorporating non-contemporaneous sequences into maximum likelihood phylogenies. *Bioinformatics* 16:395–399.

Rambaut A, Grassly NC. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput Appl Biosci.* 13:235–238.

Suchard MA, Rambaut A. 2009. Many-core algorithms for statistical phylogenetics. *Bioinformatics* 25:1370–1376.

Wertheim JO, Kosakovsky Pond SL. 2011. Purifying selection can obscure the ancient age of viral lineages. *Mol Biol Evol.* 28(12):3355–3365.

Wertheim JO, Worobey M. 2009. Dating the age of the SIV lineages that gave rise to HIV-1 and HIV-2. *PLoS Comput Biol.* 5:e1000377.

Worobey M, Gemmel M, Teuwen DE, et al. (12 co-authors). 2008. Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. *Nature* 455:661–664.